Situated Reference in a Hybrid Human-robot Interaction System

Giuliani, M., Foster, M. E., Isard, A., Matheson, C., Oberlander, J., & Knoll, A. (2010). Situated Reference in a Hybrid Human-robot Interaction System. In Proceedings of the 6th International Natural Language Generation Conference. (pp. 67-75). (INLG '10). Stroudsburg, PA, USA: Association for Computational Linguistics.

Situated Reference in a Hybrid Human-Robot Interaction System

Manuel Giuliani[1] and **Mary Ellen Foster**[2] and **Amy Isard**[3]
Colin Matheson[3] and **Jon Oberlander**[3] and **Alois Knoll**[1]
[1]Informatik VI: Robotics and Embedded Systems, Technische Universität München
[2]School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh
[3]Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh

Abstract

We present the situated reference generation module of a hybrid human-robot interaction system that collaborates with a human user in assembling target objects from a wooden toy construction set. The system contains a sub-symbolic goal inference system which is able to detect the goals and errors of humans by analysing their verbal and non-verbal behaviour. The dialogue manager and reference generation components then use situated references to explain the errors to the human users and provide solution strategies. We describe a user study comparing the results from subjects who heard constant references to those who heard references generated by an adaptive process. There was no difference in the objective results across the two groups, but the subjects in the adaptive condition gave higher subjective ratings to the robot's abilities as a conversational partner. An analysis of the objective and subjective results found that the main predictors of subjective user satisfaction were the user's performance at the assembly task and the number of times they had to ask for instructions to be repeated.

1 Introduction

When two humans jointly carry out a mutual task for which both know the plan—for example, assembling a new shelf—it frequently happens that one makes an error, and the other has to assist and to explain what the error was and how it can be solved. Humans are skilled at spotting errors committed by another, as well as errors which they made themselves. Recent neurological studies have shown that error monitoring—i.e., observing the errors made by oneself or by others—plays an important role in joint activity. For example, Bekkering et al. (2009) have demonstrated that humans show the same brain activation patterns when they make an error themselves and when they observe someone else making an error.

In this paper, we describe a human-robot interaction (HRI) system that is able both to analyse the actions and the utterances of a human partner to determine if the human made an error in the assembly plan, and to explain to the human what went wrong and what to do to solve the problem. This robot combines approaches from sub-symbolic processing and symbolic reasoning in a hybrid architecture based on that described in Foster et al. (2008b).

During the construction process, it is frequently necessary to refer to an object which is being used to assemble the finished product, choosing an unambigious reference to distinguish the object from the others available. The classic reference generation algorithm, on which most subsequent implementations are based, is the incremental algorithm of Dale and Reiter (1995), which selects a set of attributes of a target object to single it out from a set of distractor objects. In real-world tasks, the speaker and hearer often have more context in common than just the knowledge of object attributes, and several extensions have been proposed, dealing with visual and discourse salience (Kelleher and Kruijff, 2006) and the ability to produce multimodal references including actions such as pointing (van der Sluis, 2005; Kranstedt and Wachsmuth, 2005).

Foster et al. (2008a) noted another type of multimodal reference which is particularly useful in embodied, task-based contexts: *haptic-ostensive* reference, in which an object is referred to as it is being manipulated by the speaker. Manipulating an object, which must be done in any case as part of the task, also makes an object more salient and therefore affords linguistic references that in-

Figure 1: The dialogue robot

dicate the increased accessibility of the referent. This type of reference is similar to the *placing-for* actions noted by Clark (1996).

An initial approach for generating referring expressions that make use of haptic-ostensive reference was described in (Foster et al., 2009a). With this system, a study was conducted comparing the new reference strategy to the basic Dale and Reiter incremental algorithm. Naïve users reported that it was significantly easier to understand the instructions given by the robot when it used references generated by the more sophisticated algorithm. In this paper, we perform a similar experiment, but making use of a more capable human-robot interaction system and a more complete process for generating situated references.

2 Hybrid Human-Robot Dialogue System

The experiment described in this paper makes use of a hybrid human-robot dialogue system which supports multimodal human-robot collaboration on a joint construction task. The robot (Figure 1) has a pair of manipulator arms with grippers, mounted in a position to resemble human arms, and an animatronic talking head (van Breemen, 2005) capable of producing facial expressions, rigid head motion, and lip-synchronised synthesised speech. The subject and the robot work together to assemble wooden construction toys on a common workspace, coordinating their actions through speech (English or German), gestures, and facial expressions.

The robot can pick up and move objects in the workspace and perform simple assembly tasks. In the scenario considered here, both of the participants know the assembly plan and jointly execute it. The robot assists the human, explains necessary assembly steps in case the human makes an error, and offers pieces as required. The workspace is divided into two areas—one belonging to the robot and one to the human—to make joint action necessary for task success.

The system has components which use both sub-symbolic and symbolic processing. It includes a goal inference module based on dynamic neural fields (Erlhagen and Bicho, 2006; Bicho et al., 2009), which selects the robot's next actions based on the human user's actions and utterances. Given a particular assembly plan and the knowledge of which objects the user has picked up, this module can determine when the user has made an error. The system also incorporates a dialogue manager based on the TrindiKit dialogue management toolkit (Larsson and Traum, 2000), which implements the information-state based approach to dialogue management. This unique combination of abilities means that when the robot detects that its human partner has made an error—for example, picking up or requesting an assembly piece that is not needed in the current step of the building plan—it can explain to the human what the error was and what can be done to correct the mistake—for example by picking up or indicating the correct assembly piece.

Messages from all of the system's input channels (speech, object recognition, and gesture recognition) are processed and combined by a multimodal fusion component based on (Giuliani and Knoll, 2008), which is the link between the symbolic and the sub-symbolic parts of the system. The fusion component then communicates with the goal inference module, which calculates the next action instructions for the robot and also determines if the user made an error. From there, fusion combines the information from goal inference with the input data and sends unified hypotheses to the dialogue manager.

When it receives the fusion hypotheses, the dialogue manager uses the dialogue history and the physical and task context to choose a response. It then sends a high-level specification of the de-

1. System First we will build a windmill.

2. User Okay.

3. User {*picks up a yellow cube, unnecessary piece for a windmill*}

4. System You don't need a yellow cube to build a windmill.

5. System To build a windmill, you first need to build a tower.

6. System *[picking up and holding out red cube]* To build the tower, insert the green bolt through the end of this red cube and screw it into the blue cube.

7. User *[takes cube, performs action]* Okay.

Figure 2: Sample human-robot dialogue, showing adaptively-generated situated references

sired response to the output planner, which in turn sends commands to each output channel: linguistic content (including multimodal referring expressions), facial expressions and gaze behaviours of the talking head, and actions of the robot manipulators. The linguistic outputs are realised using the OpenCCG surface realiser (White, 2006).

3 Reference Generation

In this system, two strategies were implemented for generating references to objects in the world: a constant version that uses only the basic incremental algorithm (Dale and Reiter, 1995) to select properties, and an adaptive version that uses more of the physical, dialogue and task context to help select the references. The constant system can produce a definite or indefinite reference, and the most appropriate combination of attributes according to the incremental algorithm. The adaptive system also generates pronominal and deictic references, and introduces the concept of multiple types of distractor sets depending on context.

Figure 2 shows a fragment of a sample interaction in which the user picks up an incorrect piece: the robot detects the error and describes the correct assembly procedure. The underlined references show the range of output produced by the adaptive reference generation module; for the constant system, the references would all have been "the red cube". The algorithms used by the adaptive reference generation module are described below.

3.1 Reference Algorithm

The module stores a history of the referring expressions spoken by both the system and the user, and uses these together with distractor sets to select referring expressions. In this domain there are two types of objects which we need to refer to: concrete objects in the world (everything which is on the table, or in the robot's or user's hand), and objects which do not yet exist, but are in the process of being created. For non-existent objects we do not build a distractor set, but simply use the name of the object. In all other cases, we use one of three types of distractor set:

- all the pieces needed to build a target object;
- all the objects referred to since the last mention of this object; or
- all the concrete objects in the world.

The first type of set is used if the object under consideration (OUC) is a negative reference to a piece in context of the creation of a target object. In all other cases, the second type is used if the OUC has been mentioned before and the third type if it has not.

When choosing a referring expression, we first process the distractor set, comparing the properties of the OUC with the properties of all distractors. If a distractor has a different type from the OUC, it is removed from the distractor set. With all other properties, if the distractor has a different value from the OUC, it is removed from the distractor set, and the OUC's property value is added to the list of properties to use.

We then choose the type of referring expression. We first look for a previous reference (PR) to the OUC, and if one exists, determine whether it was in focus. Depending on the case, we use one of the following reference strategies.

No PR If the OUC does not yet exist or we are making a negative reference, we use an indefinite article. If the robot is holding the OUC, we use a deictic reference. If the OUC does exist and there are no distractors, we use a definite; if there are distractors we use an indefinite.

PR was focal If the PR was within the same turn, we choose a pronoun for our next reference. If it was in focus but in a previous turn, if

the robot is holding the OUC we use a deictic reference, and if the robot is not holding it, we use a pronoun.

PR was not focal If the robot is holding the OUC, we make a deictic reference. Otherwise, if the PR was a pronoun, definite, or deictic, we use a definite article. If the PR was indefinite and there are no distractors, we use a definite article, if there are distractors, we use an indefinite article.

If there are any properties in the list, and the reference chosen is not a pronoun, we add them.

3.2 Examples of the Reference Algorithm

We will illustrate the reference-selection strategy with two cases from the dialogue in Figure 2.

Utterance 4 "a yellow cube"

This object is going to be referred to in a negative context as part of a windmill under construction, so the distractor set is the set of objects needed to make a windmill: {red cube, blue cube, small slat, small slat, green bolt, red bolt}.

We select the properties to use in describing the object under consideration, processing the distractor set. We first remove all objects which do not share the same type as our object under consideration, which leaves {red cube, blue cube}. We then compare the other attributes of our new object with the remaining distractors - in this case "colour". Since neither cube shares the colour "yellow" with the target object, both are removed from the distractor set, and "yellow" is added to the list of properties to use.

There is no previous reference to this object, and since we are making a negative reference, we automatically choose an indefinite article. We therefore select the reference "a yellow cube".

Utterance 6 "it" (a green bolt)

This object has been referred to before, earlier in the same utterance, so the distractor set is all the references between the earlier one and this one—{red cube}. Since this object has a different type from the bolt we want to describe, the distractor set is now empty, and nothing is added to the list of properties to use.

There is a previous definite reference to the object in the same utterance: "the green bolt". This reference was focal, so we are free to use a pronoun if appropriate. Since the previous reference was definite, and the object being referred to does exist, we choose to use a pronoun. We therefore select the reference "it".

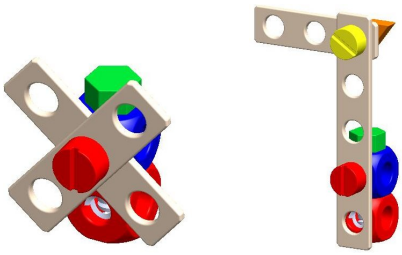
4 Experiment Design

In the context of the HRI system, a constant reference strategy is sufficient in that it makes it possible for the robot's partner to know which item is needed. On the other hand, while the varied forms produced by the more complex mechanism can increase the naturalness of the system output, they may actually be insufficient if they are not used in appropriate current circumstances—for example, "this cube" is not a particularly helpful reference if a user has no way to tell which "this" is. As a consequence, the system for generating such references must be sensitive to the current state of joint actions and—in effect—of joint attention. The difference between the two systems is a test of the adaptive version's ability to adjust expressions to pertinent circumstances. It is known that people respond well to reduced expressions like "this cube" or "it" when another person uses them appropriately (Bard et al., 2008); we need to see if the robot system can also achieve the benefits that situated reference could provide.

To address this question, the human-robot dialogue system was evaluated through a user study in which subjects interacted with the complete system. Using a between-subjects design, this study compared the two reference strategies, measuring the users' subjective reactions to the system along with their overall performance in the interaction. Based on the findings from the user evaluation described in (Foster et al., 2009a)—in which the primary effect of varying the reference strategy was on the users' subjective opinion of the robot—the main prediction for this study was as follows:

- Subjects who interact with a system using adaptive references will rate the quality of the robot's conversation more highly than the subjects who hear constant references.

We made no specific prediction regarding the effect of reference strategy on any of the objective measures: based on the results of the user evaluation mentioned above, there is no reason to expect an effect either way. Note that—as mentioned above—if the adaptive version makes incorrect choices, that may have a negative impact on users' ability to understand the system's generated references. For this reason, even a finding of



(a) Windmill (b) Railway signal

Figure 3: Target objects for the experiment

no objective difference would demonstrate that the adaptive references did not harm the users' ability to interact with the system, as long as it was accompanied by the predicted improvement in subjective judgements.

4.1 Subjects

41 subjects (33 male) took part in this experiment. The mean age of the subjects was 24.5, with a minimum of 19 and a maximum of 42. Of the subjects who indicated an area of study, the two most common areas were Mathematics (14 subjects) and Informatics (also 14 subjects). On a scale of 1 to 5, subjects gave a mean assessment of their knowledge of computers at 4.1, of speech-recognition systems at 2.0, and of human-robot systems at 1.7. Subjects were compensated for their participation in the experiment.

4.2 Scenario

This study used a between-subjects design with one independent variable: each subject interacted either with a system that used a constant strategy to generate referring expressions (19 subjects), or else with a system that used an adaptive strategy (22 subjects).[1]

Each subject built two objects in collaboration with the system, always in the same order. The first target object was the windmill (Figure 3a); after the windmill was completed, the robot and human then built a railway signal (Figure 3b). For both target objects, the user was given a building plan (on paper). To induce an error, both of the plans given to the subjects instructed them to use an incorrect piece: a yellow cube instead of a red cube for the windmill, and a long (seven-hole) slat instead of a medium (five-hole) slat for the railway signal. The subjects were told that the plan contained an error and that the robot would correct them when necessary, but did not know the nature of the error.

When the human picked up or requested an incorrect piece during the interaction, the system detected the error and explained to the human what to do in order to assemble the target object correctly. When the robot explained the error and when it handed over the pieces, it used referring expressions that were generated using the constant strategy for half of the subjects, and the adaptive strategy for the other half of the subjects.

4.3 Experimental Set-up and Procedure

The participants stood in front of the table facing the robot, equipped with a headset microphone for speech recognition. The pieces required for the target object—plus a set of additional pieces in order to make the reference task more complex—were placed on the table, using the same layout for every participant. The layout was chosen to ensure that there would be points in the interaction where the subjects had to ask the robot for building pieces from the robot's workspace, as well as situations in which the robot automatically handed over the pieces. Along with the building plan mentioned above, the subjects were given a table with the names of the pieces they could build the objects with.

4.4 Data Acquisition

At the end of a trial, the subject responded to a usability questionnaire consisting of 39 items, which fell into four main categories: Intelligence of the robot (13 items), Task ease and task success (12 items), Feelings of the user (8 items), and Conversation quality (6 items). The items on the questionnaire were based on those used in the user evaluation described in (Foster et al., 2009b), but were adapted for the scenario and research questions of the current study. The questionnaire was presented using software that let the subjects choose values between 1 and 100 with a slider. In addition to the questionnaire, the trials were also video-taped, and the system log files from all trials were kept for further analysis.

5 Results

We analysed the data resulting from this study in three different ways. First, the subjects' responses

[1] The results of an additional three subjects in the constant-reference condition could not be analysed due to technical difficulties.

Table 1: Overall usability results

	Constant	Adaptive	M-W
Intell.	79.0 (15.6)	74.9 (12.7)	$p = 0.19$, n.s.
Task	72.7 (10.4)	71.1 (8.3)	$p = 0.69$, n.s.
Feeling	66.9 (15.9)	66.8 (14.2)	$p = 0.51$, n.s.
Conv.	66.1 (13.6)	75.2 (10.7)	$p = 0.036$, sig.
Overall	72.1 (11.2)	71.8 (9.1)	$p = 0.68$, n.s.

Table 2: Objective results (all differences n.s.)

Measure	Constant	Adaptive	M-W
Duration (s.)	404.3 (62.8)	410.5 (94.6)	$p = 0.90$
Duration (turns)	29.8 (5.02)	31.2 (5.57)	$p = 0.44$
Rep requests	0.26 (0.45)	0.32 (0.78)	$p = 0.68$
Explanations	2.21 (0.63)	2.41 (0.80)	$p = 0.44$
Successful trials	1.58 (0.61)	1.55 (0.74)	$p = 0.93$

to the questionnaire items were compared to determine if there was a difference between the responses given by the two groups. A range of summary objective measures were also gathered from the log files and videos—these included the duration of the interaction measured both in seconds and in system turns, the subjects' success at building each of the target objects, the number of times that the robot had to explain the construction plan to the user, and the number of times that the users asked the system to repeat its instructions. Finally, we compared the results on the subjective and objective measures to determine which of the objective factors had the largest influence on subjective user satisfaction.

5.1 Subjective Measures

The subjects in this study gave a generally positive assessment of their interactions with the system on the questionnaire—with a mean overall satisfaction score of 72.0 out of 100—and rated the perceived intelligence of the robot particularly highly (overall mean of 76.8). Table 1 shows the mean results from the two groups of subjects for each category on the user-satisfaction questionnaire, in all cases on a scale from 0–100 (with the scores for negatively-posed questions inverted).

To test the effect of reference strategy on the usability-questionnaire responses, we performed a Mann-Whitney test comparing the distribution of responses from the two groups of subjects on the overall results, as well as on each sub-category of questions. For most categories, there was no significant difference between the responses of the two groups, with p values ranging from 0.19 to 0.69 (as shown in Table 1). The only category where a significant difference was found was on the questionnaire items that asked the subjects to assess the robot's quality as a conversational partner; for those items, the mean score from subjects who heard the adaptive references was significantly higher ($p < 0.05$) than the mean score from the subjects who heard references generated by the constant reference module. Of the six questions that were related to the conversation quality, the most significant impact was on the two questions which assessed the subjects' understanding of what they were able to do at various points during the interaction.

5.2 Objective Measures

Based on the log files and video recordings, we computed a range of objective measures. These measures were divided into three classes, based on those used in the PARADISE dialogue-system evaluation framework (Walker et al., 2000):

- Two **dialogue efficiency** measures: the mean duration of the interaction as measured both in seconds and in system turns;
- Two **dialogue quality** measures: the number of times that the robot gave explanations, and the number of times that the user asked for instructions to be repeated; and
- One **task success** measure: how many of the (two) target objects were constructed as intended (i.e., as shown in Figure 3).

For each of these measures, we tested whether the difference in reference strategy had a significant effect, again via a Mann-Whitney test. Table 2 illustrates the results on these objective measures, divided by the reference strategy.

The results from the two groups of subjects were very similar on all of these measures: on average, the experiment took 404 seconds (nearly seven minutes) to complete with the constant strategy and 410 seconds with the adaptive, the mean number of system turns was close to 30 in both cases, just over one-quarter of all subjects asked for instructions to be repeated, the robot gave just over two explanations per trial, and about three-quarters of all target objects (i.e. 1.5 out of 2) were correctly built. The Mann-Whitney test confirms that none of the differences between the two groups even came close to significance on any of the objective measures.

5.3 Comparing Objective and Subjective Measures

In the preceding sections, we presented results on a number of objective and subjective measures. While the subjects generally rated their experience of using the system positively, there was some degree of variation, most of which could not be attributed to the difference in reference strategy. Also, the results on the objective measures varied widely across the subjects, but again were not generally affected by the reference strategy. In this section, we examine the relationship between these two classes of measures in order to determine which of the objective measures had the largest effect on users' subjective reactions to the HRI system.

Being able to predict subjective user satisfaction from more easily-measured objective properties can be very useful for developers of interactive systems: in addition to making it possible to evaluate systems based on automatically available data without the need for extensive experiments with users, such a performance function can also be used in an online, incremental manner to adapt system behaviour to avoid entering a state that is likely to reduce user satisfaction (Litman and Pan, 2002), or can be used as a reward function in a reinforcement-learning scenario (Walker, 2000).

We employed the procedure used in the PARADISE evaluation framework (Walker et al., 2000) to explore the relationship between the subjective and objective factors. The PARADISE model uses stepwise multiple linear regression to predict subjective user satisfaction based on measures representing the performance dimensions of task success, dialogue quality, and dialogue efficiency, resulting in a predictor function of the following form:

$$Satisfaction = \sum_{i=1}^{n} w_i * \mathcal{N}(m_i)$$

The m_i terms represent the value of each measure, while the $\mathcal{N}$ function transforms each measure into a normal distribution using z-score normalisation. Stepwise linear regression produces coefficients (w_i) describing the relative contribution of each predictor to the user satisfaction. If a predictor does not contribute significantly, its w_i value is zero after the stepwise process.

Table 3 shows the predictor functions that were derived for each of the classes of subjective measures in this study, using all of the objective measures from Table 2 as initial factors. The R^2 column indicates the percentage of the variance in the target measure that is explained by the predictor function, while the *Significance* column gives significance values for each term in the function.

In general, the two factors with the biggest influence on user satisfaction were the number of repetition requests (which had a uniformly negative effect on user satisfaction), and the number of target objects correctly built by the user (which generally had a positive effect). Aside from the questions on user feelings, the R^2 values are generally in line with those found in previous PARADISE evaluations of other dialogue systems (Walker et al., 2000; Litman and Pan, 2002), and in fact are much higher than those found in a previous similar study (Foster et al., 2009b).

6 Discussion

The subjective responses on the relevant items from the usability questionnaire suggest that the subjects perceived the robot to be a better conversational partner if it used contextually varied, situationally-appropriate referring expressions than if it always used a baseline, constant strategy; this supports the main prediction for this study. The result also agrees with the findings of a previous study (Foster et al., 2009a)—this system did not incorporate goal inference and had a less-sophisticated reference strategy, but the main effect of changing reference strategy was also on the users' subjective opinions of the robot's interactive ability. These studies together support the current effort in the natural-language generation community to devise more sophisticated reference generation algorithms.

On the other hand, there was no significant difference between the two groups on any of the objective measures: the dialogue efficiency, dialogue quality, and task success were nearly identical across the two groups of subjects. A detailed analysis of the subjects' gaze and object-manipulation behaviour immediately after various forms of generated references from the robot also failed to find any significant differences between the various reference types. These overall results are not particularly surprising: studies of human-human dialogue in a similar joint construction task (Bard et al., In prep.) have demonstrated that the collaborators preserve quality of construction in

Table 3: PARADISE predictor functions for each category on the usability questionnaire

Measure	Function	R^2	Significance
Intelligence	$76.8 + 7.00 * \mathcal{N}(\text{Correct}) - 5.51 * \mathcal{N}(\text{Repeats})$	0.39	Correct: $p < 0.001$, Repeats: $p < 0.005$
Task	$72.4 + 3.54 * \mathcal{N}(\text{Correct}) - 3.45 * \mathcal{N}(\text{Repeats}) - 2.17 * \mathcal{N}(\text{Explain})$	0.43	Correct: $p < 0.005$, Repeats: $p < 0.01$, Explain: $p \approx 0.10$
Feeling	$66.9 - 6.54 * \mathcal{N}(\text{Repeats}) + 4.28 * \mathcal{N}(\text{Seconds})$	0.09	Repeats: $p < 0.05$, Seconds: $p \approx 0.12$
Conversation	$71.0 + 5.28 * \mathcal{N}(\text{Correct}) - 3.08 * \mathcal{N}(\text{Repeats})$	0.20	Correct: $p < 0.01$, Repeats: $p \approx 0.10$
Overall	$72.0 + 4.80 * \mathcal{N}(\text{Correct}) - 4.27 * \mathcal{N}(\text{Repeats})$	0.40	Correct: $p < 0.001$, Repeats: $p < 0.005$

all cases, though circumstances may dictate what strategies they use to do this. Combined with the subjective findings, this lack of an objective effect suggests that the references generated by the adaptive strategy were both sufficient and more natural than those generated by the constant strategy.

The analysis of the relationship between the subjective and objective measures analysis has also confirmed and extended the findings from a similar analysis (Foster et al., 2009b). In that study, the main contributors to user satisfaction were user repetition requests (negative), task success, and dialogue length (both positive). In the current study, the primary factors were similar, although dialogue length was less prominent as a factor and task success was more prominent. These findings are generally intuitive: subjects who are able to complete the joint construction task are clearly having more successful interactions than those who are not able to complete the task, while subjects who need to ask for instructions to be repeated are equally clearly not having successful interactions. The findings add evidence that, in this sort of task-based, embodied dialogue system, users enjoy the experience more when they are able to complete the task successfully and are able to understand the spoken contributions of their partner, and also suggest that designers should concentrate on these aspects of the interaction when designing the system.

7 Conclusions

We have presented the reference generation module of a hybrid human-robot interaction system that combines a goal-inference component based on sub-symbolic dynamic neural fields with a natural-language interface based on more traditional symbolic techniques. This combination of approaches results in a system that is able to work together with a human partner on a mutual construction task, interpreting its partner's verbal and non-verbal behaviour and responding appropriately to unexpected actions (errors) of the partner.

We have then described a user evaluation of this system, concentrating on the impact of different techniques for generating situated references in the context of the robot's corrective feedback. The results of this study indicate that using an adaptive strategy to generate the references significantly increases the users' opinion of the robot as a conversational partner, without having any effect on any of the other measures. This result agrees with the findings of the system evaluation described in (Foster et al., 2009a), and adds evidence that sophisticated generation techniques are able to improve users' experiences with interactive systems.

An analysis of the relationship between the objective and subjective measures found that the main contributors to user satisfaction were the users' task performance (which had a positive effect on most measures of satisfaction), and the number of times the users had to ask for instructions to be repeated (which had a generally negative effect). Again, these results agree with the findings of a previous study (Foster et al., 2009b), and also suggest priorities for designers of this type of task-based interactive system.

Acknowledgements

This research was supported by the European Commission through the JAST[2] (IST-FP6-003747-IP) and INDIGO[3] (IST-FP6-045388) projects. Thanks to Pawel Dacka and Levent Kent for help in running the experiment and analysing the data.

[2] http://www.jast-project.eu/

[3] http://www.ics.forth.gr/indigo/

References

E. G. Bard, R. Hill, and M. E. Foster. 2008. What tunes accessibility of referring expressions in task-related dialogue? In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci 2008)*. Chicago.

E. G. Bard, R. L. Hill, M. E. Foster, and M. Arai. In prep. How do we tune accessibility in joint tasks: Roles and regulations.

H. Bekkering, E.R.A. de Bruijn, R.H. Cuijpers, R. Newman-Norlund, H.T. van Schie, and R. Meulenbroek. 2009. Joint action: Neurocognitive mechanisms supporting human interaction. *Topics in Cognitive Science*, 1(2):340–352.

E. Bicho, L. Louro, N. Hipolito, and W. Erlhagen. 2009. A dynamic field approach to goal inference and error monitoring for human-robot interaction. In *Proceedings of the Symposium on "New Frontiers in Human-Robot Interaction", AISB 2009 Convention*. Heriot-Watt University Edinburgh.

H. H. Clark. 1996. *Using Language*. Cambridge University Press.

R. Dale and E. Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

W. Erlhagen and E. Bicho. 2006. The dynamic neural field approach to cognitive robotics. *Journal of Neural Engineering*, 3(3):R36–R54.

M. E. Foster, E. G. Bard, R. L. Hill, M. Guhe, J. Oberlander, and A. Knoll. 2008a. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of HRI 2008*.

M. E. Foster, M. Giuliani, A. Isard, C. Matheson, J. Oberlander, and A. Knoll. 2009a. Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *Proceedings of IJCAI-09*.

M. E. Foster, M. Giuliani, and A. Knoll. 2009b. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of ACL-IJCNLP 2009*.

M. E. Foster, M. Giuliani, T. Müller, M. Rickert, A. Knoll, W. Erlhagen, E. Bicho, N. Hipólito, and L. Louro. 2008b. Combining goal inference and natural-language dialogue for human-robot joint action. In *Proceedings of the 1st International Workshop on Combinations of Intelligent Methods and Applications at ECAI 2008*.

M. Giuliani and A. Knoll. 2008. MultiML: A general-purpose representation language for multimodal human utterances. In *Proceedings of ICMI 2008*.

J. D. Kelleher and G.-J. M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of COLING-ACL 2006*.

A. Kranstedt and I. Wachsmuth. 2005. Incremental generation of multimodal deixis referring to objects. In *Proceedings of ENLG 2005*.

S. Larsson and D. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3&4):323–340.

D. J. Litman and S. Pan. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12(2–3):111–137.

A. J. N. van Breemen. 2005. iCat: Experimenting with animabotics. In *Proceedings of AISB 2005 Creative Robotics Symposium*.

I. F. van der Sluis. 2005. *Multimodal Reference: Studies in Automatic Generation of Multimodal Referring Expressions*. Ph.D. thesis, University of Tilburg.

M. Walker, C. Kamm, and D. Litman. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3–4):363–377.

M. A. Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.

M. White. 2006. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75.